Real-time Entropic Unsupervised Violent Scenes Detection in Hollywood Movies - DYNI @ MediaEval Affect Task 2011

H. Glotin[(a,b,e)], J. Razik[(a,b)]
{glotin,razik}@univ-tln.fr
(a) Information Dynamics and Integration team (DYNI LSIS)
(e) Institut universitaire de France (IUF)

S. Paris[(a,c)]
paris@lsis.org
(c) Univ. Aix Marseille
13397 Marseille Cedex 20, France

J.-M. Prévot[(d,b)]
jmp@univ-tln.fr
(d) Computer Science Department
(b) Univ. Sud-Toulon Var
83957 La Garde Cedex, France

ABSTRACT

State of the art of the high level feature detectors, as violent scene detectors, are supervised systems. The aim of our proposition is to show that simple non supervised confidence function derived from straightforward features can perform well compared to nowdays supervised systems for this kind of hard task. Then, we develop a violent event detector independent of the kind of movies based on our previous research on basic efficient entropic movie features. We propose an entropic audiovisual confidence computed as the average of the entropies of some simple visual and acoustic features. In a first approach, we develop our system for uniform false alarm and missing costs, which is not optimal according to the official campaign criterion. However, the usual Fmeasure metrics indicates that our system is the second best among the five other -supervised- submitted systems.

Keywords

Entropy features, violent event detection, audiovisual detector, online system, unsupervised information retrieval

1. INTRODUCTION

State of the art high level feature detectors are supervised systems, including the violent event scene detectors [1]. However we assume that violent events demonstrate a specific dynamics that shall allow to compute on the fly a weak detector. Therefore, we develop an online movie violent event detector independent of the kind of movies, and with no need of labeled training dataset. Based on our previous research on basic but efficient entropic high level feature detection [2], we propose here an entropic audiovisual unsupervised confidence. It is based on the average of the entropies of some simple visual and acoustic features. This paper depicts our best official (run2), our run1 was acoustic only, and noised by acoustic stream asynchrony.

 Acknowledgment : We particularly thank Technicolor Rennes, UNIGE and IRISA TexMex for their organization of the violent scenes detection task. We thank the NII team for their visualization interface given for analyze after the official results.

In a first approach we develop our system for uniform false alarm and missing costs, which is not optimal according to the official campaign criterion. However, according to the Fmeasure criteria, our system is well performing among the five other (supervised) systems submitted to the official campaign. We think that our feature shall then allow improvements when used in supervised systems.

2. ENTROPIC VISUAL CONFIDENCE

We assume that the violent visual scene dynamics is universal and may match with fast orientation changes or other events. Thus, as in the state of the art [1], we use a line segments detector to extract orientation features. In [3] the fusion of two well known line segment detectors defined a new and fast efficient one. This operator [3] can be seen as a unified approach to statistical and structural texture analysis. An image can be represented as a single histogram computed by applying a multi-scale Local Binary Pattern [5,6] over the whole image. In very noisy images, a multiscale [4] approach is needed to obtain correct distribution. However, for a first approach, for each frame at time t, we only consider the first scale, for 12 segment lengths $\lambda_1, ..., \lambda_{12}$, and 12 orientations $\theta_1, ..., \theta_{12}$ (one every $\pi/12$). We extract one frame each second. The extraction process of this feature is then nearly twenty times faster than real-time using our toolbox [6]. In a second step, let be, for a visual frame at time t, X_t its discrete random variable with alphabet $\alpha = (\theta_i, \lambda_j), (i,j) \in [1,12]^2$, and probability mass function $p(x_t) = Pr(X_t = x_t), x_t \in \alpha$. Considering two consecutive X_t, X_{t+1}, we then propose two kinds of visual confidence. First we considered the shot average of the Kullback-Leibler distance $dKL(p(x_t), p(x_{t+1}))$, but this run has not been submitted. Second, we set $dx_t = |p(x_t) - p(x_{t+1})|$, which is normalized to estimate a probability mass function $p(dx_t)$. Then we compute its entropy $H_t = H(X_t) = -\sum_{dx_t \in \alpha} p(dx_t).log(p(dx_t))$. Finally, for each shot S, the visual confidence is set to $\gamma_v(S) = \overline{H_{t_s}}$ for each frame $t_s \in S$.

3. ENTROPIC ACOUSTIC CONFIDENCE

The extracted audio track was delayed with the visual track for some unknown reason. However we propose a simple entropic acoustic feature. First we extract Mel Filter Cepstrum Coefficients (MFCC) using the SPro toolbox [7] (window length 20 ms). We extract their speed and acceleration, and we remove the energy coefficients, yielding to 36 dimen-

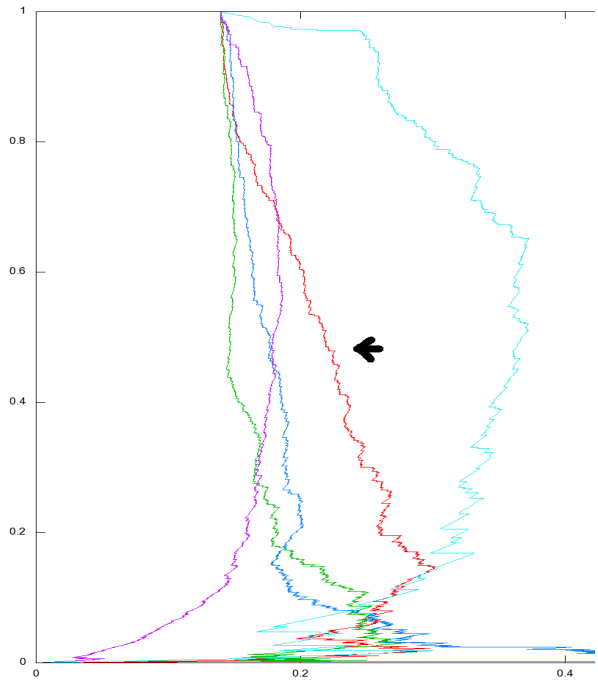

Figure 1: Official Precision (X) Recall (Y) curves of the best run of each of the six participants. The curve of our DYNI run is in red and is pointed by the arrow.

sions each 10 ms. Then we set, for each shot S the acoustic confidence $\gamma_a(S)$ to the complementary of the average of the normalized entropy of each MFCC probability distribution.

4. AUDIOVISUAL CONFIDENCE

For each shot, the audiovisual confidence to $\gamma(S) = 4 * \gamma_v(S) + \gamma_a(S)$. The overweighting of γ_v is due to the acoustic asynchrony, and its lower discriminative power that has been observed on the training set. We then threshold this final confidence in order to get twenty percent of the test set shots as positive.

5. OFFICIAL RESULTS

We remind that we did not optimized our unsupervised system according to the official weighted criterion, which is over weighting the missing cost by ten against the false alarms cost. Thus the system is not performing well according to this criterion. However the official results indicate that according to the Fmeasure, our system performs well (see Tab.1 and the Precision-Recall curve in Fig.1).

6. DISCUSSION AND CONCLUSION

The NII Lab [8] provided with their interface another analysis of this DYNI run : our 50 top list confidences over the three test movies are pointing to 14 relevant violent shots[1], which remains an interesting score.

Further work will consist first in a technical improvement : a better synchronism between the extracted audio and video tracks (nearly two seconds of delay is observed in our system between visual and acoustic streams due to the extraction system).

Second improvement shall consist in developing a more accurate audiovisual fusion. It shall be easily optimized on the training set. We shall also take into account the non uniform weighted false alarm and missing costs of the official criteria.

Considering that our unsupervised system has the second best Fmeasure, and that all the other runs are trained on 20 hours of training set, we think that our feature shall then allow improvements when used into supervised systems.

[1]These shots can be played on the NII interface pointed from http://glotin.univ-tln.fr/mediaeval2011.

Table 1: Official results table of the Fmeasure criterion of the best run of each team, at the shot level. All runs are supervised systems on nearly twenty hours of labeled movies, except our. The official criterion is given in the row "Weighted", weighting missing cost by a factor 10.

RUN name	Fmeas.	Precis.	Recall	Weighted
TECHNICOLOR,1	0,397	0,249	0,971	0,761
DYNI LSIS,2	0,293	0,242	0,372	6,470
UNIGE,4	0,289	0,178	0,774	2,838
NII,6	0,245	0,140	1,000	1,000
TUB,1	0,244	0,139	0,971	1,262
LIG,1	0,197	0,179	0,223	7,940

7. REFERENCES

[1] Demarty C.H, Penet C., Gravier G. and Soleymani M., The MediaEval 2011 Affect Task: Violent Scenes Detection in Hollywood Movies, MediaEval 2011 Workshop, Sept 2011, Pisa

[2] Glotin H., Zhao Z.Q and Ayache S., Efficient Image Concept Indexing by Harmonic and Arithmetic Profiles, in IEEE Int. Conf. on Image Proc., ICIP, Nov 2009

[3] Grompone von Gioi R., Jakubowicz J., Morel J.-M. and Randall G., LSD: A Fast Line Segment Detector with a False Detection Control, IEEE Trans. PAMI, 19, Dec 2008

[4] Paris S. and Glotin H., PyramidalMulti-Level Features for the robotVision@ICPR 2010 Challenge, ICPR 2010

[5] Paris S.,Glotin H., and Zhao Z.Q., Real-time face detection using Integral Histogram of Multi-Scale Local Binary Patterns, ICIC 2011

[6] Paris S., Scenes Objects Classification Toolbox http://www.mathworks.com/matlabcentral/fileexchange/29800-scenesobjects-classification-toolbox

[7] Gravier and al., Spro, speech signal processing toolkit, INRIA project, https://gforge.inria.fr/projects/spro

[8] Vu L., Duy-Dinh L., Shinichi S., and Duc Anh D., NII, Japan at MediaEval 2011 Violent Scenes Detection Task, in Mediaeval 2011 Proc. Demo: http://satoh-lab.ex.nii.ac.jp/users/ledduy/Demo-MediaEval/